DATA ANALYSIS I

Data for Data Mining

Types and Sources of Data

2019-20

Sources

- Bramer, M. (2013). *Principles of data mining*. Springer. [1-9]

- Witten, I. H., Frank, E. (2011). *Data Mining: Practical machine learning tools and techniques [3rd Ed.]*. Morgan Kaufmann.[3-21, 33-36]

Terms

- **Knowledge Discovery**: Non-trivial extraction of implicit, previously unknown and potentially useful information from data.

- **Methodology**: Obtaining the data – Preprocessing – Data mining – Interpretation.

- **Data Mining**: Finding and describing structural patterns in data.

- **Machine Learning**: It is used to extract information from the raw data and provides the technical basis of data mining.

- In general we have a set of *examples* (*objects, instances*), each of which comprises the values of a number of *variables* (*features, attributes*). In data mining, the set of objects is called *dataset*.

Describing Structural Patterns

- Data mining algorithms produce an output in the form of rules or some other kind of patterns.

- **Example**

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Patterns

- Part of a structural description of the weather data

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true    then play = no
If outlook = overcast                  then play = yes
If humidity = normal                   then play = yes
If none of the above                   then play = yes
```

Applications of Data Mining

- **Classification**: It involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories known as *classes*. In *classification learning*, the learning scheme is presented with a set of classified objects from which it is expected to learn a way of classifying unseen objects.

- **Numeric prediction**: It is a variant of classification learning in which the outcome is a numeric value rather than a class.

- **Associations**: Rules that strongly associate different attribute values. In *association learning*, any association among attributes is sought, not just ones that predict a particular class value.

- **Clustering**: Clustering algorithms examine data to find groups of objects that are similar.

Labeled Data and Supervised Learning

- **Labeled data**: There is a specially designated attribute and the aim is to use the data given to predict the value of that attribute for instances that have not yet been seen.

- **Supervised learning**: It is data mining using labeled data. It contains classification and numeric prediction.

- **Example (Classification)**: A hospital may want to classify medical patients into those who are at high, medium or low risk of acquiring a certain illness.

- **Example (Numeric prediction)**: A company wishes to predict a numeric value, such as a company's profits or a share price.

Unlabeled Data and Unsupervised Learning

- **Unlabeled data**: Data that does not have any specially designated attribute.

- **Unsupervised learning**: It is data mining using unlabeled data. The aim is to extract the most information we can from the data available. It contains association rules and clustering.

- **Example (Association rules)**: If we know the purchases made by all the customers at a store for say a week, we may be able to find relationships that will help the store market its products more effectively in the future.

- **Example (Clustering)**: An insurance company might group customers according to income, age, types of policy purchased or prior claims experience.

Ethics, Using Personal Information

- **Data about people**: The use of data for data mining has serious ethical implications.

- **Example**: When applied to people, data mining is frequently used to discriminate; who gets the loan, who gets the special offer... Certain kinds of discrimination (racial, sexual, religious…) are not only unethical but also illegal.

- **Using Personal Information**: It is necessary to determine the conditions under which the data was collected and for what purposes it may be used.

- **Anonymization**: It is a process of removing personally identifiable information from datasets.

- **Example (Reidentification)**: Over 85% of Americans can be identified from publicly available records using just three pieces of information: five-digit zip code, birth date (including year), and sex. Over half of Americans can be identified from just city, birth date, and sex.